Conditional Simulation Discussion Topics:

Following are a series of discussion topics that EPA is interested in pursuing with CPG in the upcoming meeting on February 17th. To facilitate productive discussions, EPA asks that CPG bring live GIS files including data group boundaries, simulated surfaces and a selection of diagnostic data. It is not expected that these data would be exhaustive, but rather illustrative of the work that has been conducted and suitable for CPG to guide EPA through answers to some of the discussion topics below. EPA may at a later date request copies of simulation output and diagnostics to the extent practical.

1) The EPA is interested in how data groupings in the conditional simulation model may bias simulation results. In particular, there are a relatively large number of groups some of which are supported by a relatively small number of sample locations. Simulating independent realizations within these relatively small groups may artificially constrain the range of concentrations that could be expected within them. Rather than simulating each group independently, a model based on spatially varying means, or external drift is a more common approach to handle stratification of the population based on covariate information (Goovaerts, 1997). This alternative approach is a more holistic analysis than that proposed by CPG. EPA is interested in a discussion of these two alternative approaches.

   The basic form of the CPG simulation approach is as follows:

   The CPG simulation approach recognizes that with a strongly varying mean, the statistical properties of the concentrations may vary substantially among groups, represented here as the *Trend* component of the equation. The CPG approach is to assume that within each data group, the mean and variance are unrelated to neighboring groups, and that each group can be simulated independently from neighboring groups. Effectively the CPG approach simulates the *Trend* component directly but independently within each data group.

   The spatially varying mean approach proceeds by first subtracting the Trend component (i.e. detrending, based on data group means) and then treating the resulting residual (*Error term)* as a single spatially varying random function. With this approach simulated concentrations vary smoothly between data groups, but still have relatively steep gradients along borders as opposed to the sharper breaks that would be induced by the CPG approach. The means of 100 replicate maps would be very similar among these approaches, however, the spatially varying mean approach allows for the possibility of mixing higher and lower concentration areas to blur across data group boundaries.

a. EPA anticipates that projections of benefit may differ between the two approaches because the CPG approach constrains simulated distributions within data groups more strongly than would the spatially varying means approach. Because there are no validation data it is not currently known which approach provides the more accurate assessment.
b. Recognizing that models used for the FS will have considerable uncertainty, EPA suggests discussions focus on the pros and cons of a potentially more or less optimistic simulation approaches and how varying degrees of optimism may drive the nature of design sampling programs.

2) Related to item 1 above there is also interest in flexibility in the size and number of groups and statistical support for group separation. Large numbers of groups may present logistics issues in the future as there may be desire to rapidly update the simulation models. It may be advantageous from this perspective to reduce the number of groups based on consideration of statistical separation of groups.
    a. It may be helpful to run a simulation with a very small number of groups to see if answers to key questions change substantively.
    b. May want to consider a more formal statistical procedure of estimating group means and variances. One method which considers the relative strength of differences between group means and the number of samples is based on a Bayesian technique which evaluates complete and partial pooling of data across groups. With this method, the mean per group is composed of a weighted average of the group mean and the overall pooled mean where weights are a function of group sample size and variance. This approach is described by Gelman and Hill (pg. 252, 2007). EPA does not necessarily advocate this approach, but suggests that it may provide a method for deriving group means and variances in a way that is less sensitive to individual data groups and also takes account of the relative sample sizes per group. This approach may also provide a streamlined way to update group means as design data become available.
3) EPA would like to discuss appropriate simulation diagnostics necessary for inclusion in reports documenting analyses. These diagnostics should be adequate to document that the computer programs are set up and executing properly, and further that the software settings are appropriate to reproduce important characteristics of the sample concentration data. At a minimum EPA will be interested in evaluations of how concentration maps should match three properties of the sample data;
    a. simulated surfaces should match actual concentrations at the data locations,
    b. the semivariogram of simulated surfaces should match the selected semivariogram model used to generate the simulation, and
    c. the histogram of the simulated values should match the declustered histogram of the sample data.

    EPA asks that CPG have available diagnostic plots and data to support discussions about what CPG has done thus far and what EPA may want to have available for future evaluation.

4) Mean concentrations in areas that are identified as having had no deposition are generally sparsely sampled. Key calculations may be sensitive to uncertainty in estimated mean TCDD

concentrations in these areas. EPA views CPGs assertion that these areas have low contaminant concentrations as a plausible, but as yet, unvalidated hypothesis. In the recent discussions of this simulation approach there was discussion of design sampling being focused on areas of high uncertainty, which EPA generally agrees is a sensible approach. However, EPA also suggests that upcoming discussions should also focus on the potential need for confirmation and or design sampling to test hypotheses underlying the construction of the simulation model.

   a. This would include validation sampling within these non-depositional areas.
   b. Other parts of the model may also need additional testing, such as sampling designed specifically for estimation of short scale variation (i.e. nugget effect).

5) Nugget effect
   a. EPA will provide CPG with data supporting nugget effect calculations presented in previous memoranda.
   b. EPA suggests that it may be helpful to run simulations with and without nugget to test sensitivity of simulation output to this parameter.
6) EPA anticipates extensive discussion of how best to use simulated contaminant maps
   a. to inform various questions related to the FS,
   b. How to simulate remedial effectiveness for fate and transport model
   c. How to simulate benefit of remedy for SWAC vs RAL relationship.
   d. Consideration of potential ongoing application during optimization of remedial design and implementation.
7) Potential application to depth of contamination at later stages in the project.

References

Gelman, A. and J.Hill. 2007. *Data Analysis Using Regression and Multilevel/Hierarchical Models.* Analytical Methods for Social Research. Cambridge University Press.

Goovaerts, P. 1997. *Geostatisitcs for Natural Resources Evaluation.* Applied Geostatistics Series. Oxford University Press.